A Corpus of Word-Aligned Asked and Anticipated Questions in a Virtual Patient Dialogue System

Ajda Gokcen, Evan Jaffe, Johnsey Erdmann, Michael White, Douglas Danforth
The Ohio State University
Department of Linguistics, Department of Family Medicine
{gokcen.2,jaffe.59,erdmann.10}@osu.edu, mwhite@ling.osu.edu
doug.danforth@osumc.edu

Abstract

We present a corpus of virtual patient dialogues to which we have added manually annotated gold standard word alignments. Since each question asked by a medical student in the dialogues is mapped to a canonical, anticipated version of the question, the corpus implicitly defines a large set of paraphrase (and non-paraphrase) pairs. We also present a novel process for selecting the most useful data to annotate with word alignments and for ensuring consistent paraphrase status decisions. In support of this process, we have enhanced the earlier Edinburgh alignment tool (Cohn et al., 2008) and revised and extended the Edinburgh guidelines, in particular adding guidance intended to ensure that the word alignments are consistent with the overall paraphrase status decision. The finished corpus and the enhanced alignment tool are made freely available.

Keywords: monolingual word alignment, paraphrase detection, virtual patient dialogues

1. Introduction

Many systems for detecting paraphrases or measuring semantic similarity rely on alignments of words and phrases (whether explicitly or implicitly). However, to our knowledge only two monolingual corpora with manually annotated gold standard alignments have been developed to date, and none in a naturalistic task setting: the MSRP (Brockett, 2007) and Edinburgh (Cohn et al., 2008) corpora.[1] Moreover, both of these corpora have additional shortcomings. In particular, the Edinburgh corpus consists only of aligned paraphrase pairs, rather than containing both paraphrase and non-paraphrase pairs; and although the MSRP corpus does include a mix of paraphrase and non-paraphrase pairs, the word alignments were annotated without taking into consideration the paraphrase status of the sentence pair. This potentially makes the MSRP alignments less useful than they might otherwise be, given that Xu et al. (2014) have recently shown that there can be considerable benefit to modeling word alignment and paraphrase classification as a joint process.

In this paper, we present a corpus of 104 dialogues between early stage medical students and a virtual standardized patient (Figure 1) from which we have manually annotated word alignments for 942 sentence pairs. With this dialogue system (Danforth et al., 2013), the interpretation task is to determine for each question asked by the medical student which of the set of questions anticipated by the content author best matches the student's question. In previous work on developing a log-linear ranking model for this system (Jaffe et al., 2015), we found the interpretation task to be quite challenging, since it requires discriminating among many topically related questions; by contrast, the MSRP task (Dolan et al., 2004) is to classify pairs of sentences as paraphrases or non-paraphrases, where the pairs themselves are mostly unrelated. In particular, we found that alignments produced by Meteor (Denkowski and Lavie, 2011) were less helpful than anticipated, given that Meteor played an important role in an ensemble of MT metrics (Madnani et al., 2012) that until recently yielded the best paraphrase classification results on the MSRP corpus. As such, we expect our corpus to be a valuable asset for research on improved methods of monolingual alignment (Thadani et al., 2012; Yao et al., 2013) as well as paraphrase detection and semantic similarity measurement more generally.[2]

In constructing the corpus, we have made use of a novel process for selecting data to annotate with both word alignments and paraphrase status. We have also enhanced the Edinburgh alignment tool to better support this process, and revised and extended the Edinburgh guidelines. In the ensuing sections, we review this process and these tool and guideline enhancements. Together with the paper, we are making the corpus of dialogues and gold standard word alignments freely available, along with the enhanced alignment tool.

2. Annotation Process

The corpus consists of two main parts — the dialogues, and the word-level alignments. The dialogues are a set of 104 conversations of 5347 total turns, where each turn consists of an asked question, a label, and a response.

[1] See Xu et al. (2014) for a review of paraphrase corpora without gold standard alignments. Also closely related is the much larger WikiAnswers Paraphrase Dataset (Fader et al., 2013), which again does not have manually annotated alignments and which exhibits much less topical coherence.

[2] Naturally, we do not discourage the development of additional annotated corpora for enabling research on monolingual alignment and paraphrase detection, e.g. ones from naturalistic task settings primarily involving declarative sentences rather than questions.

Input	Gold Label	ChatScript System Response
hello there, why did you come in to see us today	what brings you in today	i was hoping you could help me with my back pain, it really hurts! it has been awful. i can barely even tie my shoes
it sounds like that has a big impact on your life, can you tell me more about the pain	tell me more about your back pain	my buddy lives on the second floor in an apartment building. we needed to bring everything down and there were only 3 of us...
could you rate the pain	rate the pain	[pain2] right now, i would call my pain a 7. when i take the ibuprofen, it goes down to about a 5 or 6
is the pain constant	is the pain constant	[fear2] it is pretty constant although sometimes it is a little better or worse...
how is this affecting your life	how is this affecting you	[fear1]i haven't been able to work, i am a little worried about the bills
what do you do for work	what do you do for work	i am an auto mechanic. i have a garage a couple of blocks from here...

Table 1: Example dialogue turns with labels

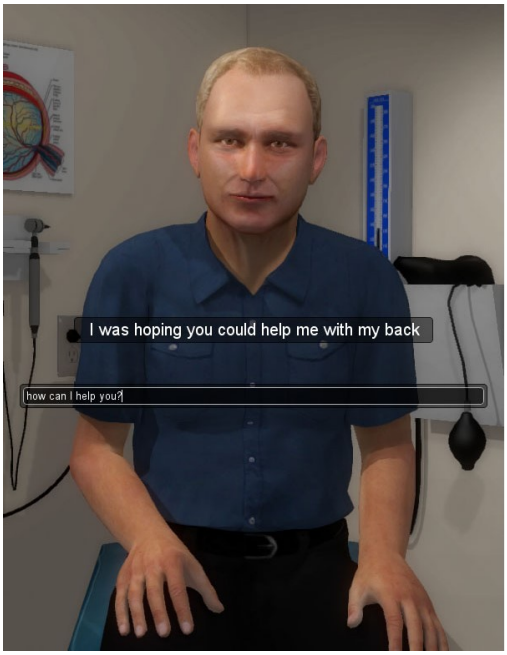


Figure 1: Example exam room and virtual patient avatar

A label is defined as the canonical form of a question asked to the virtual patient. For example, *What brings you in today?* might be the label for a number of variants of that label, such as, *Why are you here today?*, *Can you tell me why you've come in today?*, or *What seems to be the problem?*.

The gold labels in the data used in our previous machine learning experiments (Jaffe et al., 2015) came from a hand-crafted pattern-matching system called ChatScript (Wilcox, 2011), whose output was then corrected manually. These labels were written by content authors as part of the ChatScript system to recognize variants and match them to their canonical form, or label.

A similar process was used for the current, larger set of dialogues, labeling questions using ChatScript output and some hand-correction by content authors. Additionally, annotators sometimes found cases where automatic labels were not consistent with the human paraphrase judgments. In these cases, the labels were manually corrected to reflect the human paraphrase status.

Paraphrases are defined in the dialogues as those pairs of asked questions having the same gold label. Given the labels, question pairs can be generated for alignment annotation as either paraphrases or non-paraphrases by comparing labels.

In order to maximize the utility of a limited amount of human annotation, question pairs are selected by choosing the ones expected to be the most informative first. Inspired by active learning (Cohn et al., 1994), the most useful data points are expected to be those closest to a classifier's decision boundaries. Automatic alignment scores are used as a cheap estimate of where boundaries would potentially occur in a classifier, where well-aligned non-paraphrases and poorly-aligned paraphrases would be close to one another on opposite sides of a categorical boundary.

For each question, we collect the three paraphrase pairs with the lowest automatic alignment and the three non-paraphrase pairs with the highest alignment; these six pairs constitute a batch. Questions for which there do not exist at least three paraphrases and at least three non-paraphrases in the training dialogues are excluded. It is worth noting that this strategy for choosing difficult cases first means that annotation will likely speed up over time, as the relatively harder alignment cases are exhausted.

Additionally, in order to get good coverage across labels, the ordered batches are grouped by label and then a batch is taken from each label before taking the next most informative batch from the same label. This way, we ensure getting batches from many labels and that for any given label, we choose the most informative batches first.

This process of selecting batches draws attention to border cases, which are also those most likely to benefit from a human judgment of paraphrase status (potentially requiring label correction). The alignment tool makes it easy to check dialogue context and see all questions with the same label (question variants), which allows annotators to disagree with paraphrase status and identify questions whose label should be hand-corrected. For example, annotators saw the question pair *are you in pain currently* and *has the pain caused you to miss work*, which was considered a paraphrase because both questions were grouped together with variants such as *has your back pain made you take days off of work*. Annotators suspected that the pair was not a para-

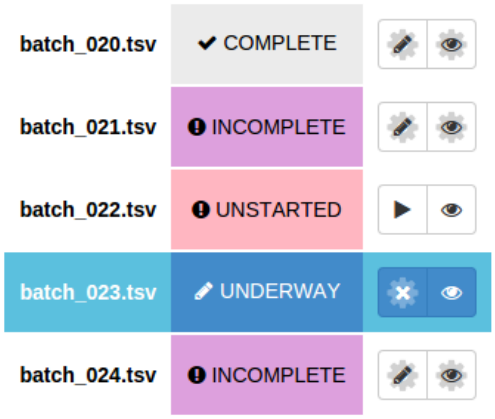


Figure 2: A logged-in user's view of the batch navigation, with one batch currently in progress.

(a) Paraphrase judgment

(b) Context links

Figure 3: Two of the new features of the tool, allowing for more nuanced handling of non-paraphrase alignment.

phrase, and using the label variants, quickly concluded that *are you in pain currently* was mislabeled.

The final corrected set of labels (and question variants with each label) was compiled by the content creators using a spreadsheet listing all question variants, their labels, and instances in the final annotations where the paraphrase judgment between the given question and any other question differed from the paraphrase judgment in the initial corpus. Changes made to the labels of the question variants were informed by the principle that all questions with a given label ought to be paraphrases with one another and ought not be paraphrases with questions under any other label. The one exception to this principle was for the special "negative symptoms" label, which serves as a catch-all for rarely asked questions regarding irrelevant symptoms (e.g. *Do you have any rashes?* or *Have you noticed any swelling?*).

3. Tool Enhancements

Very few tools accommodate human annotation of alignment and paraphrase data, let alone allow for an arbitration process that yields gold standard annotations. Prior to this work, the tool that came closest was the descendant of the tool used to develop the Edinburgh corpus,[3] made specifically for crowdsourcing via Mechanical Turk. It rests on the assumption that sentence pairs to be aligned are paraphrases, and the two-step process comprises annotation followed by single-user grading. Our process required several innovations: a private interface for a smaller set of annotators, a means of straightforward cross-user arbitration, and readily viewable links to contextual information as well as a mechanism for paraphrase judgments in order to account for non-paraphrases.

GoldAlign, the interface we created to fit this role, manages users and versions in addition to providing a graphical interface with all the needed input fields. Users log in as either single annotators or as arbitrators of multiple such annotators. They select a dataset to work with, which corresponds to a directory containing any number of variable-length data files referred to as *batches*, and can immediately see an overview of their completion of the batches in the dataset, as in Figure 2. The dataset-batch structure simply allows for the division of the data into convenient subsets. A full description of the data and file structure will be released along with the freely available tool.[4]

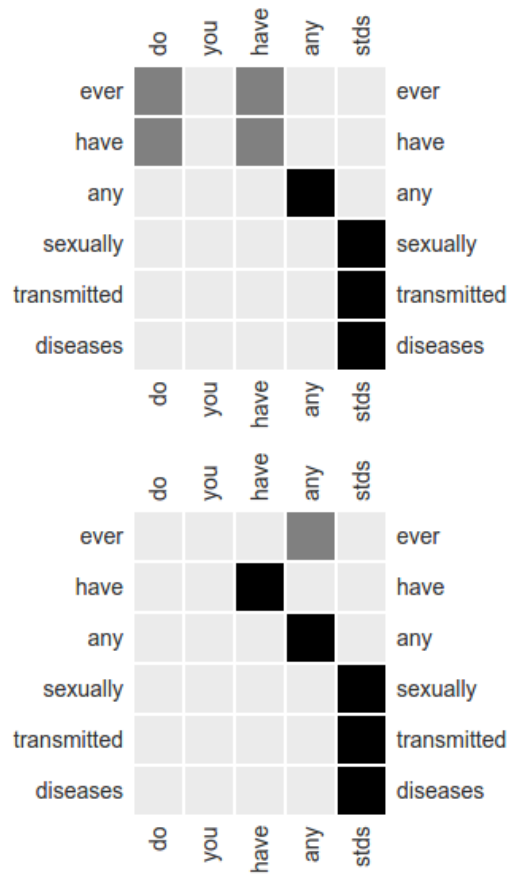
The grid-based alignment workspace is largely the same in both annotation and arbitration modes. For each sentence pair in a batch, users see an initial alignment grid in which squares corresponding to word-level alignments may be given a "sure" or "possible" value (colored black and gray, respectively).

Departing from previous tools in which it is given that any two sentences being compared are paraphrases, users are asked to give a boolean paraphrase judgment (that is, whether or not the candidate sentences are paraphrases of one another). In addition, users are given the ability to write free-form comments should they have an issue with the data, such as an error in the corpus or personal confusion regarding the best annotation. Finally, users are given links to files containing the contexts of both candidate sentences, so that the alignment and paraphrasing decisions may be contextually-informed as needed (Figure 3).

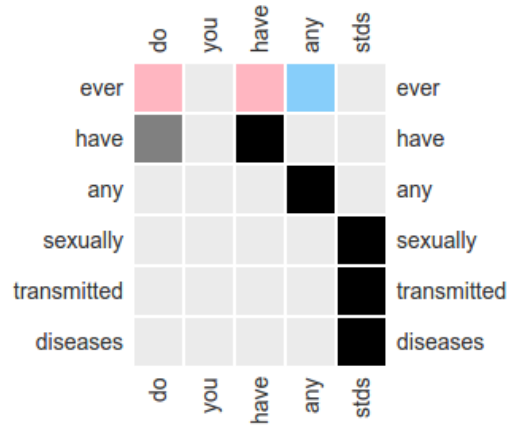
If the user is performing an initial annotation task, the grid is populated by the calculations of an automated aligner such as Meteor, whereas the grid in arbitration mode shows both of the original annotators' selections, color-coded to bring attention to any discrepancies between them so that the arbiter may decide on the official annotation. In neither case does the user modify either the initial alignment or any other user's alignments. The annotations of the two users being arbitrated are merely shown to help the arbitrating user determine their own finalized annotations.

[3] We thank Chris Callison-Burch for making this tool available.

[4] The most up-to-date version of GoldAlign can also be found at `https://github.com/ajdagokcen/goldalign-repo`.



(a) Two users' respective alignment annotations for a given sentence pair.



(b) The arbitration grid for comparing and finalizing the two users' alignment annotations in (a). Black and gray squares are included in the final annotation file, while other colors show the original users' annotations without affecting the final version.

Figure 4: The alignment grids for a pair of sentences as seen in each user's annotation mode and in color-coded arbitration mode between the two of them.

A comparison between the two grid views is shown in Figure 4. The colored squares in the arbitration grid indicate that both of the original two annotations included possible alignments that were not included in the arbitrating user's final annotation.

Annotators		Precision		Recall		F1	
		Sure	Poss	Sure	Poss	Sure	Poss
M	G	0.16	0.45	0.36	0.12	0.22	0.19
A	G	0.66	0.71	0.80	0.76	0.72	0.74
J	G	0.59	0.73	0.68	0.58	0.63	0.64
A	M	0.41	0.12	0.22	0.46	0.29	0.19
J	M	0.49	0.18	0.25	0.53	0.33	0.27
A	J	0.48	0.67	0.46	0.50	0.47	0.57

Table 2: Inter-annotator agreement over distinct words for gold arbitrated alignments (G), human annotators (A and J) and Meteor (M)

Annotators		Precision		Recall		F1	
		Sure	Poss	Sure	Poss	Sure	Poss
M	G	0.20	0.30	0.60	0.29	0.30	0.40
A	G	0.76	0.74	0.87	0.75	0.81	0.74
J	G	0.61	0.60	0.79	0.60	0.69	0.60
A	M	0.64	0.31	0.25	0.34	0.36	0.33
J	M	0.77	0.44	0.34	0.46	0.48	0.45
A	J	0.57	0.55	0.64	0.54	0.60	0.55

Table 3: Inter-annotator agreement over distinct atomic phrase pairs for gold arbitrated alignments (G), human annotators (A and J) and Meteor (M)

4. Corpus Statistics

Batches are taken from a corpus that has 104 dialogues with a total of 5437 user turns. Each dialogue has 52 turns on average (min. 3, max. 141, s.d. 25.1). Each user turn consists of an average of 7 words (min. 1, max. 76, s.d. 5.1). There are 39,073 total words and 290 unique labels. Each label has an average of 19 turns (min. 1, max. 486, s.d. 38.1).

Excluding the 486 turns with the special "negative symptoms" label, which is a catch-all for questions regarding irrelevant symptoms, there are 4951 user turns. Each dialogue has 48 turns on average (min. 3, max. 112, s.d. 21.4). Each user turn consists of an average of 7 words (min. 1, max. 76, s.d. 5.2). There are 36,847 total words and 289 unique labels. Each label has an average of 17 turns (min. 1, max. 122, s.d. 26.4).

There are 157 annotated and arbitrated batches, representing 942 sentence pairs, with 441 paraphrases and 495 non-paraphrases.

Using the scripts for calculating inter-annotator agreement distributed with the Edinburgh corpus, Tables 2 and 3 show word- and phrase-based agreement between gold arbitrated alignments G and human annotators A and J as well as automatic Meteor alignments, M. As the tables show, agreement for the human annotators is much higher than for Meteor.

5. Revised and Extended Guidelines

We took the Edinburgh alignment annotation guidelines as our starting point, revising and extending them as necessary to clarify difficult decisions that arose as our two annotators went through a trial set of 47 batches of question pairs. Many of the additional guidelines were needed to handle non-paraphrases, which are not included in the Edinburgh corpus. In developing our guidelines, we took the earlier MSRP alignment guidelines into account where they were consistent with the Edinburgh ones. However, contrary to the MSRP guidelines—where annotators were not told whether two sentences were supposed to be in an entailment relationship—we considered it essential for the alignment decisions to be consistent with the overall decision as to whether the two questions were taken to be paraphrases in the dialogue context. In particular, we decided that if two questions are taken to be paraphrases, they must have at least one content word aligned.

We have grouped our 16 additional guidelines into 7 categories, as listed below. The revised guidelines, along with the original Edinburgh guidelines, are distributed with the corpus and alignment tool.

Paraphrase Status

- Sentences that are judged to be paraphrases should have at least one content word aligned (either sure or possible); if it's not possible to align any content words, then the sentences should not be considered paraphrases.

Multiple Occurrences of a Word or Phrase

- When a word or phrase appears twice in a question, align the one in most direct correspondence in terms of syntax and word order and mark the other one as possible.

Time

- With time reference, use possible alignments when different event times are implied, not just when different tenses are used: for example, present progressive and present perfect should be annotated with possible alignments.

Direct Substitution

- Use sure alignments when phrases are directly substitutable (in both directions) despite differences in syntax (e.g. *for* and *to help with*).

- Use sure alignments with hypernyms if they are fully substitutable in context, i.e. no important meaning is lost (e.g. *hi* for *good afternoon*).

Words

- Don't align function words (e.g. articles, WH-question words, and NPIs) that are modifying unaligned content words. Possessive pronouns should be aligned if they have the same reference.

- Use possible alignments when a pronoun is aligned to a full definite noun phrase (e.g. *it* for *the pain*).

- Prepositions that are only serving a syntactic function but don't affect the meaning should be left unaligned.

Verb Clusters

- When all the auxiliaries in a verb cluster are the same in the source and target, they should be singly aligned rather than block aligned (consistent with the Edinburgh corpus, though this detail is not spelled out in their guidelines).

- When a verb cluster in the source and target are of the same length and the main verb is inflected in the same way, they should be singly aligned rather than block aligned (e.g. singly align *will be priced* with *would be priced*, as in the Edinburgh corpus, though again this detail is not spelled out in their guidelines, and it's unclear how consistently such cases are handled).

- The verb *have* with a condition, such as *have nausea*, should be treated as a light/support verb construction and be block aligned with its equivalent in paraphrases (e.g. with *be nauseous*). Conversely, in non-paraphrases, *have* should not be aligned if the *have* + condition construction does not have an equivalent, even when there is an identical form of *have* (e.g. don't align *have nausea* with *have STDs* at all).

- Align verb clusters with adjectives or adverbs that provide the same time reference using possible alignments (e.g. *have had* for *previous*) when there is no corresponding modifier for the adjective or adverb (e.g. just align *in the past* with *previous* if such a modifier is present).

- Number agreement should be treated as a minor syntactic divergence and therefore possible alignments should be used (e.g. with the verbs *make* and *makes*).

- Don't align words that describe different things, including verbs that describe different events (even if they're the same vague verb like *doing*). A specific property or location of nouns (e.g., *pain in legs* vs. *pain in back*) is not necessarily enough to make them different in this way.

- Prepositions that do affect meaning like phrasal verbs should be aligned.

Tokenization and Spelling Errors

- If there's a typo that affects the tokenization (e.g. a run-on error), then it needs to be fixed before the annotation can be done. The tokenization error should

be noted in the comments. Otherwise, if the intended word should be aligned, align the typo as a possible alignment.

6. Conclusion

In this paper, we have presented a corpus of virtual patient dialogues to which we have added manually annotated gold standard word alignments. Since each asked question in the dialogues is assigned a hand-corrected label indicating the anticipated question it best corresponds to, the corpus implicitly defines a large set of paraphrase (and non-paraphrase) pairs. We have also presented a novel process for selecting data to annotate with word alignments and ensuring consistent paraphrase status decisions. In support of this process, we have enhanced the Edinburgh alignment tool (Cohn et al., 2008) and revised and extended the Edinburgh guidelines, in particular adding guidance intended to ensure that the word alignments are consistent with the overall paraphrase status decision. The finished corpus and the enhanced alignment tool are made freely available.

Acknowledgments

We would like to acknowledge Kellen Maicher who created the virtual environment and Bruce Wilcox who authored ChatScript and customized the software for this project. We also acknowledge the expert technical assistance of Laura Zimmerman who managed the laboratory and organized student involvement in this project.

The work reported in this paper was made possible by a Targeted Investment in Excellence grant from the Department of Linguistics at the Ohio State University. The material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This project was additionally funded (in part) by a National Board of Medical Examiners (NBME) Edward J. Stemmler, MD Medical Education Research Fund grant (NBME 1112-064). The project does not necessarily reflect NBME policy, and NBME support provides no official endorsement. This project was also supported by funding from the Department of Health and Human Services Health Resources and Services Administration (HRSA D56HP020687).

References

Brockett, C. (2007). Aligning the 2006 RTE corpus. In *Technical Report MSR-TR-2007-77, Microsoft Research*.

Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. *Mach. Learn.*, 15(2):201–221, May.

Cohn, T., Callison-Burch, C., and Lapata, M. (2008). Constructing corpora for development and evaluation of paraphrase systems. *Computational Lingustics*, 34(4):597–614.

Danforth, D. R., Price, A., Maicher, K., Post, D., Liston, B., Clinchot, D., Ledford, C., Way, D., and Cronau, H. (2013). Can virtual standardized patients be used to assess communication skills in medical students? In *Proceedings of the 17th Annual IAMSE Meeting*, St. Andrews, Scotland.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*, pages 350–356, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Fader, A., Zettlemoyer, L., and Etzioni, O. (2013). Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Jaffe, E., White, M., Schuler, W., Fosler-Lussier, E., Rosenfeld, A., and Danforth, D. (2015). Interpreting questions with a log-linear ranking model in a virtual patient dialogue system. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 86–96, Denver, Colorado, June. Association for Computational Linguistics.

Madnani, N., Tetreault, J., and Chodorow, M. (2012). Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190, Montréal, Canada, June. Association for Computational Linguistics.

Thadani, K., Martin, S., and White, M. (2012). A joint phrasal and dependency model for paraphrase alignment. In *Proceedings of COLING 2012: Posters*, pages 1229–1238, Mumbai, India, December. The COLING 2012 Organizing Committee.

Wilcox, B. (2011). Chatscript. `http://chatscript.sourceforge.net/`.

Xu, W., Ritter, A., Callison-Burch, C., Dolan, W. B., and Ji, Y. (2014). Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics (TACL)*, 2(1).

Yao, X., Van Durme, B., Callison-Burch, C., and Clark, P. (2013). A lightweight and high performance monolingual word aligner. In *Proceedings of ACL short*.